NATIONAL UNIVERSITY OF SINGAPORE

Mathematics PhD Qualifying Exam Paper 4

Stochastic Processes and Machine Learning

(January 2022)

Time allowed : 3 hours

INSTRUCTIONS TO CANDIDATES

1. Please write your matriculation/student number only. Do not write your name.
2. Including this page, the examination paper comprises **4** printed pages.
3. This examination contains **5** questions. Answer all of them. **Properly justify** your answers.
4. There is a total of **ONE HUNDRED (100)** points. The points for each question are indicated at the beginning of the question.
5. Please start each part of a question (i.e., (a), (b), etc.) on a new page. Answer all parts of a question together.
6. This is an OPEN BOOK exam. No electronic device (such as calculator, tablet, laptop or phone) is allowed. You need to have your reference materials in hard copy with you.
7. A list containing information on the probability density / mass function, mean, variance and moment generating functions of some common distributions has been provided on the other side for possible consultation.

Q1 (24 points) Let $\mathbb{A}$ be a $d \times d$ non-negative definite matrix with maximum eigenvalue $\lambda_{\max}(\mathbb{A})$. Let $\{\mathbf{u}_i\}_{i=1}^{n+1}$ be i.i.d. random vectors sampled uniformly at random from the unit sphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$. Consider the random variable

$$\mathcal{Q}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{u}_i^\top \mathbb{A}\, \mathbf{u}_{i+1}.$$

- **(a) (12 points)** Calculate $\mathrm{Var}[\mathcal{Q}_n]$ in terms of n, d and $(\mathbb{A}_{ij})_{1 \le i,j \le d}$.
- **(b) (12 points)** Show that, for $x > 0$, we have

$$\mathbb{P}[|\mathcal{Q}_n| \ge x] \le 2 \exp\left(- \frac{x^2}{2\lambda_{\max}(\mathbb{A})^2} \right).$$

Q2 (18 points) Let $\{X_i\}_{i \ge 1}$ be i.i.d. exponential random variables with parameter $\lambda > 0$, and let N be an independent Poisson random variable with parameter μ. Let $Y = \sum_{i=1}^{N} X_i$.

- **(a) (6 points)** Calculate the moment generating function of Y, clearly explaining when is finite and when it is infinite.
- **(b) (12 points)** Show that, for $x > \mu/\lambda$, we have

$$\mathbb{P}[Y \ge x] \le \exp\left(-(\sqrt{\lambda x} - \sqrt{\mu})^2 \right).$$

Q3 (8 points) Consider a sequence of independent tosses of a fair coin. Calculate the expected number of coin tosses necessary to obtain the first occurrence of two consecutive Heads.

Q4 (**20 points**) Consider a discrete-time Markov decision process with finite state space $\mathcal{S}$ and action space $\mathcal{A}$. We use the usual notation of S_t, A_t, R_t to denote the state, action and reward at time t respectively. Let the transition probability kernel be $p(s', r|s, a) = P[S_{t+1} = s', R_{t+1} = r | S_t = s, A_t = a]$.

- (a) (3 points) Write down the definition of the value function v_π with respect to a policy π. You may assume that we consider a discount rate of $0 < \gamma < 1$ when computing the returns.
- (b) (3 points) Write down the Bellman's optimality equation that the value function corresponding to an optimal policy should satisfy.
- (c) (7 points) Show that there exists a unique solution to the Bellman's optimality equation.
- (d) (7 points) Is an optimal policy always unique ? If so, prove this statement. If not, give a counterexample.

Q5 (30 points)

Residual Neural Networks (ResNet) are slightly different from fully connected neural networks. For an input $x \in \mathbb{R}$, a simple ResNet is given by

$$f(x) = x + \sum_{i=1}^{n} v_i \sigma(w_i\, x),$$

where $x, w_i \in \mathbb{R}$, $v_i \in \mathbb{R}$, and $\sigma(z) = \text{ReLU}(z) = \max(z, 0)$ is the ReLU activation function. We assume that the weights w_i's are iid standard Gaussian random variables, i.e. $w_i \sim \mathcal{N}(0, 1)$. Similarly, we assume that v_i's are iid Gaussian random variables with variance $1/n$, i.e. $v_i \sim \mathcal{N}(0, 1/n)$. For the sake of simplification, we assume that the input x is a standard Gaussian variable, $x \sim \mathcal{N}(0, 1)$, independent from the weights w and v.

- (a) (6 points) Let $x \in \mathbb{R}^d$. What is $\mathbb{E}[f(x)]$ and $\text{Var}[f(x)]$? (the expectation $\mathbb{E}$ is taken with respect to random variables x, w_i's, and v_i's)
- (b) (9 points) Is $f(x)$ Gaussian? What is the distribution of $f(x)$ in the limit $n \to \infty$ (only a rigorous justification will be accepted)?
- (c) (10 points) We would like to understand what happens when we increase the number of layers. Consider the following ResNet of depth $L \geq 1$,

  $$\begin{aligned} y_0(x) &= x, \\ y_\ell(x) &= y_{\ell-1}(x) + \sum_{i=1}^{n} v_{i,\ell}\, \sigma(w_{i,\ell}\, y_{\ell-1}(x)), \quad 1 \leq \ell \leq L. \end{aligned}$$

  where $x \sim \mathcal{N}(0, 1), w_{i,\ell} \sim \mathcal{N}(0, 1)$, $v_{i,\ell} \sim \mathcal{N}(0, 1/n)$, and all these random variables are independent.

  Let $\ell \in \{1, \ldots, L\}$. Express $\mathbb{E}[y_\ell(x)]$ and $\text{Var}[y_\ell(x)]$ as a function of ℓ (the expectation $\mathbb{E}$ is taken with respect to random variables x, $w_{i,k}$, and $v_{i,k}$, for all i, k).
  Give an interpretation to the result.
- (d) (5 points) Let $\ell \in \{2, \ldots, L\}$. Does $y_\ell(x)$ converge (weakly) to a Gaussian distribution in the limit $n \to \infty$? (justify).

—— End of Paper ——

- Bernoulli (p) :
  $\mathbb{P}(X=i)=\begin{cases} p \text{ if } i=1 \\ 1-p \text{ if } i=0. \end{cases}$
  $\mathbb{E}[X]=p, \quad \text{Var}[X]=p(1-p), \quad \mathbb{E}[e^{tX}]=(1-p)+pe^t.$

- Binomial (n,p):
  $\mathbb{P}(X=i)=\binom{n}{i}p^i(1-p)^{n-i}; 0 \leq i \leq n.$
  $\mathbb{E}[X]=np, \quad \text{Var}[X]=np(1-p), \quad \mathbb{E}[e^{tX}]=[(1-p)+pe^t]^n.$

- Geometric (p) :
  $\mathbb{P}(X=i)=(1-p)^{i-1}p; i \geq 1.$
  $\mathbb{E}[X]=\frac{1}{p}, \quad \text{Var}[X]=\frac{1-p}{p^2}, \quad \mathbb{E}[e^{tX}]=\frac{pe^t}{1-(1-p)e^t}$ for $t < -\log(1-p)$.

- Poisson (λ):
  $\mathbb{P}(X=i)=e^{-\lambda}\frac{\lambda^i}{i!}; i \geq 1.$
  $\mathbb{E}[X]=\lambda, \quad \text{Var}[X]=\lambda, \quad \mathbb{E}[e^{tX}]=\exp(\lambda(e^t-1)).$

- Uniform (a,b) :
  $f(x)=\begin{cases} \frac{1}{b-a} \text{ if } a \leq x \leq b \\ 0 \text{ otherwise .} \end{cases}$
  $\mathbb{E}[X]=(a+b)/2, \quad \text{Var}[X]=\frac{(b-a)^2}{12}, \quad \mathbb{E}[e^{tX}]=\frac{e^{tb}-e^{ta}}{t(b-a)}$ if $t \neq 0$.

- Uniform on the square $(a,b)\times(c,d)$:
  $f(x,y)=\begin{cases} \frac{1}{(b-a)(d-c)} \text{ if } a \leq x \leq b, c \leq y \leq d \\ 0 \text{ otherwise .} \end{cases}$

- Normal / Gaussian ($N(\mu,\sigma^2)$):
  $f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2}).$
  $\mathbb{E}[X]=\mu, \quad \text{Var}[X]=\sigma^2, \quad \mathbb{E}[e^{tX}]=\exp(\mu t+\frac{1}{2}\sigma^2 t^2).$

- Exponential (λ):
  $f(x)=\begin{cases} \lambda\exp(-\lambda x) \text{ if } x>0 \\ 0 \text{ otherwise.} \end{cases}$
  $\mathbb{E}[X]=1/\lambda, \quad \text{Var}[X]=1/\lambda^2, \quad \mathbb{E}[e^{tX}]=\frac{\lambda}{\lambda-t}$ for $t<\lambda$.